

Learning 3D Deformation of Animals from 2D Images

Angjoo Kanazawa¹, Shahar Kovalsky², Ronen Basri² and David Jacobs¹

¹University of Maryland College Park ²Weizmann Institute of Science

Abstract

Understanding how an animal can deform and articulate is essential for a realistic modification of its 3D model. In this paper, we show that such information can be learned from user-clicked 2D images and a template 3D model of the target animal. We present a volumetric deformation framework that produces a set of new 3D models by deforming a template 3D model according to a set of user-clicked images. Our framework is based on a novel locally-bounded deformation energy, where every local region has its own stiffness value that bounds how much distortion is allowed at that location. We jointly learn the local stiffness bounds as we deform the template 3D mesh to match each user-clicked image. We show that this seemingly complex task can be solved as a sequence of convex optimization problems. We demonstrate the effectiveness of our approach on cats and horses, which are highly deformable and articulated animals. Our framework produces new 3D models of animals that are significantly more plausible than methods without learned stiffness.

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Geometric algorithms, languages, and systems

1. Introduction

Recent advances in computer vision and graphics have enabled the collection of high-quality 3D models with tools such as multi-view stereo [FP10] and commercial depth-sensors [IKH*11]. However, it is still difficult to obtain models of highly articulated and deformable objects like animals. Today, searching Turbosquid for “chair” returns 24,929 results, while “cat” returns only 164 results. On the other hand, the Internet is filled with images of cats. The goal of our work is to create new 3D models of animals by modifying a template 3D model according to a set of user-clicked 2D images. The user clicks serve as positional constraints that guide the shape modification.

In order to modify the shape realistically, we argue that it is critical to understand how an animal can deform and articulate. For example, looking at many images of cats shows that a cat’s body may curl up like a ball or twist and that its limbs articulate, but its skull stays mostly rigid. Hence, when modifying a cat 3D model, we should restrict the amount of deformation allowed around the skull, but allow larger freedom around limb joints and the torso.

In this work, we propose a novel deformation framework that aims to capture an animal-specific 3D deformation model from a set of annotated 2D images and a template 3D model. Our framework is inspired by the idea of *local stiffness*, which specifies the amount of distortion allowed for a local region. Stiffness is used in 3D surface deformation methods to model natural bending at joints and elastic deformations [PJS06, BPGK06]. In previous methods, the stiffness is provided by users or learned from a set of vertex-

aligned 3D meshes in various poses [PJS06]. Instead, we learn stiffness from user-clicked 2D images by imposing sparsity; the idea is that large distortion is only allowed for those regions that require high deformation across many images. To our knowledge, our work is the first to learn stiffness of a 3D model from annotated 2D images.

Figure 1 shows an overview of our proposed framework. Given a stock 3D cat mesh and target images of cats, a user provides 3D-to-2D point correspondences by clicking key features in images. These are passed on to the proposed algorithm, which simultaneously deforms the mesh to fit each cat’s pose and learns a cat-specific model of 3D deformation. In the end, we obtain new 3D models for each target image and a stiffness model that describes how cats may deform and articulate.

Contributions: Our primary contribution is a deformation framework that learns an animal-specific model of local stiffness as it deforms the template model to match the user-clicked 2D-to-3D correspondences. Specifically,

1. We propose a locally bounded volumetric deformation energy that controls the maximal amount of distortion applied to local regions of the model using the recent optimization techniques of [KABL14]. The bounds act as a local stiffness model of the animal, which we learn by imposing a L1 sparsity penalty. The final deformation is orientation preserving and has worst-case distortion guarantees.
2. We show that both the deformation and the stiffness bounds can

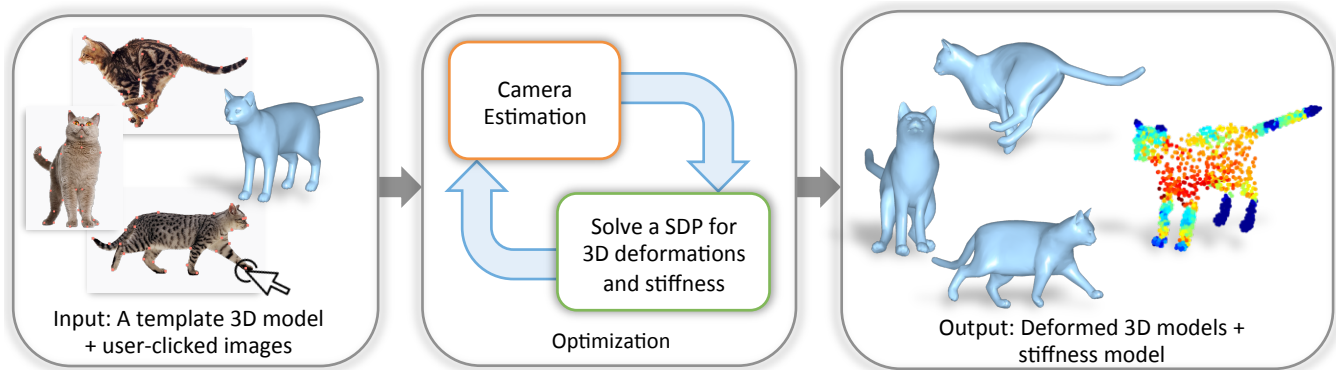


Figure 1: Overview. Our inputs are a template 3D model of an animal and a set of images with user clicked 3D-to-2D point correspondences. The algorithm then alternates between solving for the camera viewpoint and the 3D deformations for all images. Our novel formulation allows us to solve for the deformation for each image and the stiffness model of the animal jointly in a single semidefinite program (SDP). The outputs of our algorithm are a set of deformed 3D models and the stiffness model, which specifies the rigidity of every local region of the animal (red indicates high deformability and blue indicates rigidity).

- be solved jointly as a sequence of convex optimization problems.
3. We demonstrate the effectiveness of our framework on cats and horses, which are challenging animals as they exhibit large degrees of deformation and articulation.

2. Related Work

Works that modify a template 3D model to fit images can be roughly divided into two categories: those that are class-agnostic, and those that have a class-specific deformation prior learned from data or provided by users. Methods that do not use any class-specific prior make use of strong image cues such as silhouettes [VdA13, hTg10] or contour drawings [KSvdP09]. These approaches focus on fitting a single 3D model into a single image, while we focus on learning a class-specific prior as we modify the template 3D model to fit multiple images. Recently, Kholgade et al. introduced an exciting new photo editing tool that allows users to perform 3D manipulation by aligning 3D stock models to 2D images [KSES14]. Our approach complements this application, which is only demonstrated for rigid objects.

More closely related to our approach are works that make use of prior knowledge on how the 3D model can change its shape. Many works assume a prior is provided by users or artists in the form of kinematic skeletons [GWBB09, AB15, TSR*14a, BTG*12, YY14, TSR*14b] or painted stiffness [PJS06]. Since obtaining such priors from users is expensive, many methods learn deformation models automatically from data [BV99, AKP*04, ASK*05, CKC10, dATTS08, LD14, SY07]. Anguelov et al. [AKP*04] use a set of registered 3D range scans of human bodies in a variety of configurations to construct skeletons using graphical models. Blanz and Vetter [BV99] learn a morphable model of human faces from 3D scans, where a 3D face is described by a linear combination of basis faces. Given a rough initial alignment, they fit the learned morphable models to images by restricting the model to the space spanned by the learned basis. Similarly [ASK*05, HSS*09] learn a statistical model of human bodies from a set of 3D scans. Popa et

al. [PJS06] learn the material stiffness of animal meshes by analyzing a set of vertex-aligned 3D meshes in various poses.

One of the biggest drawbacks in learning from 3D data is that it requires a large set of registered 3D models or scans, which is considerably more challenging to obtain compared to a set of user-clicked photographs. All of these methods rely on 3D data with the exception of Cashman et al. [CF13]. They learn a morphable model of non-rigid objects such as dolphins from annotated 2D images and a template 3D model. Our work is complementary to their approach in that they focus on intra-class shape variation such as fat vs thin dolphins, while we focus on deformations and articulations due to pose changes. The use of a morphable model also makes their approach not suitable for objects undergoing large articulations.

Using 2D images requires camera parameters for projecting the deformed 3D models to image coordinates. Cashman et al. [CF13] assume a rough camera initialization is provided by a user, but we estimate the camera parameters directly from user-clicked 2D-to-3D correspondences. There are many works regarding the estimation of camera parameters from image correspondences, and their discussion is outside the scope of this paper. We refer the reader to [HZ04] for more details.

There is a rich variety of mesh deformation techniques in the literature [BS08, ZHS*05, BPGK06, SSP07, SA07]. The main idea is to minimize some form of deformation objective that governs the way the mesh is modified according to user-supplied positional constraints. Common objectives are minimization of the elastic energy [TPBF87] or preservation of local differential properties [LSCO*04]. The solution can be constrained to lie in the space of natural deformations, which are learned from exemplar meshes [SP04, SZGP05, DSP06, MTTG11, PJS06]. Our approach is related to these methods, except that we learn the space of deformations from a set of annotated 2D images. [BS08] offers an excellent survey on linear surface deformation methods. While simple and efficient to use, surface deformation methods suffer from unnatural volumetric changes for large deformations [ZHS*05, BPGK06].

Our work is based on a volumetric representation, which we discuss in detail in the next section.

3. Problem statement and background

We consider the problem of modifying a template 3D mesh of an animal according to a set of user-clicked photographs of the target animal. Our goal is to produce plausible 3D models guided by the annotated images, not necessarily obtaining precise 3D reconstructions of the images. In particular, given a sparse set of 2D-to-3D correspondences obtained from user-clicks, we wish to solve for a set of class-specific 3D deformations that faithfully fit the image annotations.

More formally, we are given a 3D template model, represented by a surface mesh $\mathbf{S} \subset \mathbb{R}^3$ as well as N images of class instances I^1, \dots, I^N . Each image is associated with a sparse set of user prescribed point correspondences to the 3D model; namely, the i 'th image I^i comes with pairs $\{(\mathbf{x}_k^i, \mathbf{p}_k^i)\}$ relating the surface point $\mathbf{x}_k^i \in \mathbf{S}$ to a 2D image location $\mathbf{p}_k^i \in \mathbb{R}^2$. Our goal is to leverage the N annotated images to learn a deformation model \mathcal{D} capturing the possible deformations and articulations of the object class. In particular, for each image I^i we wish to find a deformation $\Phi^i \in \mathcal{D}$ that maps its 3D landmark points $\{\mathbf{x}_k^i\}$ to their corresponding image points $\{\mathbf{p}_k^i\}$ once projected to the image plane; namely, satisfying

$$\begin{bmatrix} \mathbf{p}_k^i \\ 1 \end{bmatrix} = \Pi^i \begin{bmatrix} \Phi^i(\mathbf{x}_k^i) \\ 1 \end{bmatrix}, \quad (1)$$

where $\Pi^i \in \mathbb{R}^{3 \times 4}$ is the camera projection matrix for the i 'th image. In what follows we assume weak perspective projection, which is an orthographic projection followed by scaling of the x and y coordinates:

$$\Pi = \begin{bmatrix} \alpha_x & & & \\ & \alpha_y & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 & t_1 \\ \mathbf{r}_2 & t_2 \\ \mathbf{0} & 1 \end{bmatrix}. \quad (2)$$

\mathbf{r}_1 and \mathbf{r}_2 are the first two rows of the object rotation matrix, t_1, t_2 are the first two coordinates of the object translation, and $\frac{\alpha_x}{\alpha_y}$ specifies the camera aspect ratio. Its parameters can be solved in a least squares approach given six or more 3D-to-2D point correspondences. Please see [HZ04] for more information. Note that perspective projection may be similarly handled.

3.1. Parameterized deformation model

We parameterize the deformations of the surface model \mathbf{S} by introducing an auxiliary tetrahedral mesh enclosed within the surface, $\mathbf{M} = (\mathbf{V}, \mathbf{T})$, where $\mathbf{V} \in \mathbb{R}^{3 \times n}$ is a matrix of n coarse vertex coordinates and $\mathbf{T} = \{t_j\}_{j=1}^m$ is a set of m tetrahedra (tets). Every surface point $\mathbf{x} \in \mathbf{S}$ can then be written as a linear combination of the vertices \mathbf{V} . In particular, for the landmark points we set $\mathbf{x}_k^i = \mathbf{V}\boldsymbol{\alpha}_k^i$, where $\boldsymbol{\alpha}_k^i \in \mathbb{R}^n$ is a coefficient vector computed by linear moving least squares [Lev98]. Figure 2 shows the surface and the tetrahedral mesh of a template cat model. The use of a tetrahedral mesh introduces a notion of volume to the model making it more robust at preserving volumetric detail [dAST*08, ZHS*05].

Deformations of \mathbf{M} thereby induce deformations of the surface \mathbf{S} . Specifically, we shall consider continuous piece-wise linear

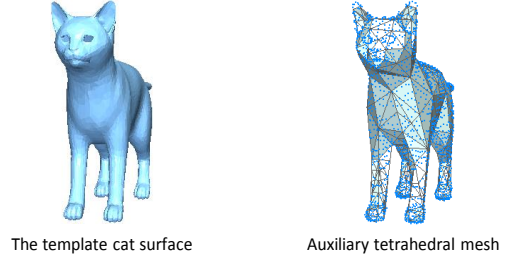


Figure 2: A template 3D surface and its auxiliary tetrahedral mesh with surface vertices shown in blue dots.

(CPL) maps $\Phi : \mathbf{M} \rightarrow \mathbb{R}^3$, whereby the deformation, restricted to the j 'th tet, is defined by the affine map $\mathbf{v} \mapsto A_j \mathbf{v} + \mathbf{t}_j$. Φ maps the vertices \mathbf{V} to new locations $\mathbf{U} \in \mathbb{R}^{3 \times n}$. In fact, Φ is uniquely determined by the new vertex locations \mathbf{U} ; for the j 'th tet, the following full rank linear system holds

$$(\mathbf{u}_{j1} \ \mathbf{u}_{j2} \ \mathbf{u}_{j3} \ \mathbf{u}_{j4}) = [A_j \ \mathbf{t}_j] \begin{pmatrix} \mathbf{v}_{j1} & \mathbf{v}_{j2} & \mathbf{v}_{j3} & \mathbf{v}_{j4} \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad (3)$$

where \mathbf{v}_j . and \mathbf{u}_j . are its four vertices in the original and the deformed mesh respectively.

We denote by $A_j = A_j(\mathbf{U})$ the linear part of each affine transformation, linearly expressed in terms of the new vertex locations \mathbf{U} . Lastly, note that subject to a deformation $\Phi = \Phi_{\mathbf{U}}$ the location of the landmark points can be simply expressed as $\hat{\mathbf{x}}_k^i = \Phi_{\mathbf{U}}(\mathbf{x}_k^i) = \mathbf{U}\boldsymbol{\alpha}_k^i$. This relationship along with the positional constraints are depicted in Figure 3.

3.2. Landmark-guided 3D deformation

Our goal is to deform the template \mathbf{S} such that (1) is satisfied without introducing local distortions to its shape. A popular approach to prevent distortion is minimizing the as-rigid-as-possible (ARAP) functional [ACOL00, SA07]:

$$f_{\text{ARAP}}(\mathbf{U}) = \sum_{j=1}^m \|A_j - R_j\|_F^2 |t_j|, \quad (4)$$

where $R_j \in SO(3)$ is the closest rotation to A_j and $|t_j|$ is the normalized volume of the j 'th tet. Intuitively, ARAP tries to keep the local transformations applied to each tet of the mesh as similar as possible to a rigid transformation. Note that while the ARAP functional is non-convex, it is convex-quadratic for fixed rotations R_j .

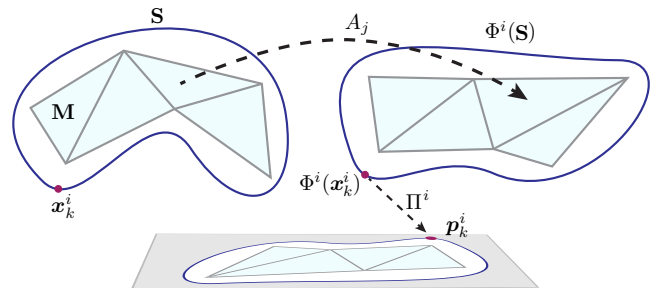


Figure 3: Illustration of the deformation model.

The ARAP functional minimizes the ℓ_2 -norm of a “non-rigidity” measure, which strives to evenly distribute local deviations from rigid transformation. As such, it fails to faithfully represent articulation and local deformations. Moreover, it is not straightforward to adapt this functional alone to benefit from having many annotated image exemplars. In this work, we also use the ARAP functional, but allow non-uniform distribution of distortion by assigning local stiffness as described in the next section.

4. Learning stiffness for articulation and deformation

Natural objects do not usually deform in a uniform manner; some parts such as joints deform a lot more while parts such as the limbs and skull stay rigid. In order to model such deformation and articulation, we introduce the notion of local stiffness, which specifies how much distortion is allowed at each tet. We learn local stiffness from data using a sparsity promoting energy, so large deformations are concentrated in regions that require them across many images.

We depart from the traditional skeleton models, which are a set of rigid sticks connected by deformable joints [ASK*05, YP08]. While skeletons excel at modeling articulation, they only possess two level of stiffness, rigid or not. In contrast, our model can represent multiple levels of stiffness, which is essential for representing local deformations. Moreover, using local stiffness does not require prior knowledge, such as the number of sticks and joints. In this section we discuss how we simultaneously deform the template \mathbf{S} to match each of the images I_1, \dots, I_N while learning the stiffness.

4.1. Modeling local stiffness

Denote by \mathbf{U}^i the deformation mapping \mathbf{S} to the i 'th image I^i , and by $\{A_j^i\}$ the linear transformations associated with its tets. Inspired by [Lip12, KABL14], we control deformations by explicitly imposing constraints on their linear parts.

First we require that each A_j^i satisfies

$$\det(A_j^i) \geq 0, \quad (5)$$

which entails that the mapping is locally injective and orientation preserving; in particular, tets may not flip. Second, we bound the *local isometric distortion* with the constraint

$$\max \left\{ \|A_j^i\|_2, \|A_j^{i-1}\|_2 \right\} \leq 1 + \varepsilon + s_j \quad (6)$$

where $\|\cdot\|_2$ is the operator (spectral) norm. The small constant $\varepsilon \geq 0$ is common for all tets and governs the degree of global non-rigidity. $s_j \geq 0$ is the local *stiffness* for the j 'th tet controlling how much this particular tet may deform. Note that ε and s_j are not image specific (i.e. they are independent of i) and encode the class-prior of how an object class can deform and articulate.

Intuitively, $\|A_j^i\|_2$ and $\|A_j^{i-1}\|_2$ quantify the largest change of Euclidean length induced by applying A_j^i to any vector. Therefore, Equation (6) bounds local length changes by a factor of $1 + \varepsilon + s_j$. If, for example, $\varepsilon = s_j = 0$ then A_j^i must be a rotation; looser bounds allow “less locally isometric” deformations. In practice, ε is set to a small value and is fixed throughout the experiments.

4.2. Optimizing articulation and deformation

Subject to these constraints, we propose minimizing an energy comprising three terms:

$$f = f_{\text{DEFORM}} + \lambda f_{\text{POS}} + \eta f_{\text{STIFFNESS}}. \quad (7)$$

f_{DEFORM} is defined via the ARAP deformation energy (4) as

$$f_{\text{DEFORM}} = \frac{1}{N} \sum_{i=1}^N f_{\text{ARAP}}(\mathbf{U}^i). \quad (8)$$

f_{POS} is defined by

$$f_{\text{POS}} = \frac{1}{N} \sum_{i=1}^N \sum_k \left\| \begin{bmatrix} \mathbf{p}_k^i \\ 1 \end{bmatrix} - \Pi^i \begin{bmatrix} \mathbf{U}^i \boldsymbol{\alpha}_k \\ 1 \end{bmatrix} \right\|_2^2, \quad (9)$$

which accounts for the user prescribed correspondences and the camera parameters, aiming to satisfy (1). Lastly, we set

$$f_{\text{STIFFNESS}} = \|\mathbf{s}\|_1, \quad (10)$$

where \mathbf{s} is the vector whose elements are the local stiffness bounds $\{s_j\}$. This L1 regularization encourages most s_j to be 0, so that only those tets that must distort are allowed to do so.

λ is a parameter that controls the trade-off between satisfying the constraints and preserving the original shape of \mathbf{M} . η is a parameter that controls the strength of the stiffness regularization. As η increases, it forces most A_j to stay rigid and as η approaches 0 the solution approaches that of the ARAP functional and the positional constraints. See Section 5 for parameter settings.

In conclusion, jointly deforming the template \mathbf{S} to match each of the images I_1, \dots, I_N , while estimating the local stiffness boils down to the following optimization problem:

$$\begin{aligned} \min_{\{\mathbf{U}^i\}, \{\Pi^i\}, \mathbf{s}} \quad & f_{\text{DEFORM}} + \lambda f_{\text{POS}} + \eta f_{\text{STIFFNESS}} \quad (11) \\ \text{s.t.} \quad & A_j^i = A_j^i(\mathbf{U}^i), \quad \forall j = 1, \dots, m, i = 1, \dots, N \\ & \det(A_j^i) \geq 0, \\ & \max \left\{ \|A_j^i\|_2, \|A_j^{i-1}\|_2 \right\} \leq 1 + \varepsilon + s_j, \\ & s_j \geq 0. \end{aligned}$$

Note that usually in prior work, deformations are solved independently for each set of positional constraints, since there is nothing that ties multiple problems together. Introducing a shared stiffness field allows us to leverage information from multiple images and improve the quality of results for all images.

4.3. Realizing the optimization

Optimizing (11) is not straightforward, as it involves the non-convex constraint (6). We realize these constraints in a convex optimization framework based on the construction presented in [KABL14] for optimization subject to bounds on the extremal singular values of matrices.

This previous work makes the observation that the set of matrices whose maximal singular value, σ_{\max} , is bounded from above by

some constant $\Gamma \geq 0$ is convex and can be written as a linear matrix inequality (LMI):

$$\mathcal{C}^\Gamma = \left\{ A \in \mathbb{R}^{n \times n} : \begin{pmatrix} \Gamma I & A \\ A^T & \Gamma I \end{pmatrix} \succeq 0 \right\}. \quad (12)$$

It is further shown that for any rotation matrix $R \in SO(n)$, the set

$$RC_\gamma = \left\{ RA \in \mathbb{R}^{n \times n} \mid \frac{A + A^T}{2} \preceq \gamma I \right\}, \quad (13)$$

is a maximal convex subset of the non-convex set of matrices with non-negative determinant whose minimal singular value, σ_{\min} , is bounded from below by some constant $\gamma \geq 0$. This calls for an iterative algorithm in which R is updated in each iteration so as to explore the entire set of matrices with bounded minimum singular value. As suggested by [KABL14], a natural choice for R is the closest rotation to A . This choice, in turn, also minimizes the ARAP functional in Equation (4) for a fixed A .

In order to employ the convex optimization framework of [KABL14], we rewrite the constraints (5) and (6) as

$$1/c_j \leq \sigma_{\min}(A_j^i) \leq \sigma_{\max}(A_j^i) \leq c_j \text{ and } \det(A_j^i) \geq 0,$$

with $c_j = 1 + \varepsilon + s_j$. This follows by observing that $\|A_j^i\|_2 = \sigma_{\max}(A_j^i)$ and $\|A_j^{i-1}\|_2 = 1/\sigma_{\min}(A_j^i)$. Plugging (11) into the framework of [KABL14] then yields the following optimization problem:

$$\begin{aligned} \min \quad & f_{\text{DEFORM}} + \lambda f_{\text{POS}} + \eta f_{\text{STIFFNESS}} \quad (14) \\ \text{s.t.} \quad & A_j^i = A_j^i(\mathbf{U}^i), \quad \forall j = 1, \dots, m, i = 1, \dots, N \\ & A_j^i \in \mathcal{C}^{\Gamma_j^i}, \\ & A_j^i \in R_j^i \mathcal{C}_{\gamma_j^i}, \\ & s_j \geq 0, \\ & \Gamma_j^i \leq (1 + \varepsilon + s_j), \\ & \frac{1}{(1 + \varepsilon + s_j)} \leq \gamma_j^i, \end{aligned}$$

whose optimization variables are $\{\mathbf{U}^i\}, \{\Gamma_j^i\}, \{\gamma_j^i\}$ and \mathbf{s} .

Lastly, we note that the last constraint of (14) is convex; in fact, following a standard derivation (e.g., see [BV04]), it can be equivalently rewritten as the convex second-order cone constraint

$$\sqrt{4 + (\gamma_j - (1 + \varepsilon + s_j))^2} \leq \gamma_j + (1 + \varepsilon + s_j).$$

Therefore, with fixed $\{R_j^i\}$ and $\{\Pi^i\}$, Equation (14) is a semidefinite program (SDP) and can be readily solved using any SDP solver. However, note that the entire problem is not convex due to the interaction between R_j^i, \mathbf{U}^i , and Π^i . Thus, we take a block-coordinate descent approach where we alternate between two steps: (a) update R_j^i and Π^i fixing \mathbf{U}^i , (b) update \mathbf{U}^i fixing R_j^i and Π^i via solving Equation (14). As in [KABL14], we find that allowing the surface to deform gradually makes the algorithm less susceptible to local minima. To this end, we initialize the procedure with a large η , which controls the degree of non-rigidity, and slowly reduce its value as the algorithm converges. This algorithm is outlined in Algorithm 1.

Algorithm 1: Jointly solving for the deformations and the stiffness

Input: Template 3D mesh \mathbf{S} , its auxiliary tetrahedral mesh $\mathbf{M} = (\mathbf{V}, \mathbf{T})$, and N 3D-to-2D annotated images $\{I^i\}$
Output: N deformed auxiliary tetrahedral meshes vertices $\{\mathbf{U}^i\}$, the projection matrices $\{\Pi^i\}$, and the stiffness model \mathbf{s}

maxIter = 10;
 $\tilde{\mathbf{U}}^i = \mathbf{V}$, $i = 1 \dots N$; // initialize
for $\eta \leftarrow \eta_{\max}$ to η_{\min} do // warm start
 $\mathbf{U}^{i(0)} = \tilde{\mathbf{U}}^i$;
 $t = 0$;
 repeat
 Compute $\Pi^{i(t)}$ by solving Equation (1) with $\mathbf{U}^{i(t)}$;
 Compute the polar decompositions $A_j^{i(t)} = R_j^{i(t)} S_j^{i(t)}$;
 Update $\{\mathbf{U}^{i(t+1)}\}, \mathbf{s}^{(t+1)}$ by solving Equation (14) with $\Pi^{i(t)}$ and $R_j^{i(t)}$;
 $t = t + 1$;
 until convergence or $t > \text{maxIter}$
 $\tilde{\mathbf{U}}^i = \mathbf{U}^{i(t)}$;
return $\{\mathbf{U}^{i(t)}\}, \{\Pi^{i(t)}\}, \mathbf{s}^{(t)}$

5. Experimental Detail

We use our approach as described to modify a template 3D mesh according to the user-clicked object pose in 2D images. We first compare our approach with the recent method of Cashman et al. [CF13], which is the closest work to ours with publicly available source code [CF]. We then present an ablation study where key components of our model are removed in order to evaluate their importance and provide qualitative and quantitative evaluations.

We experiment with two object categories, cats and horses. We collected 10 cat and 11 horse images in a wide variety of poses from the Internet. Both of the template 3D meshes were obtained from the Non-rigid World dataset [BBK07]. These templates consist of ~ 3000 vertices and ~ 6000 faces, which are simplified and converted into tetrahedral meshes of 510, 590 vertices and 1500, 1700 tets for the cat and the horse respectively via a tet generation software [Si15]. We manually simplify the mesh in MeshLab until there are around 300 vertices. We found automatic simplification methods over-simplify skinny regions and fine details, leading to a poor volumetric tet-representation. The cat template and its auxiliary tetrahedral mesh are shown in Figure 2. The template mesh used for horses can be seen in Figure 7. For all experiments we set $\varepsilon = 0.01$, and $\lambda = 10$. In order to allow gradually increasing levels of deformation, we use $\eta_{\max} = 0.5$ and $\eta_{\min} = 0.05$ with 10 log-steps in between for all experiments. The values for η and λ were set by hand, but deciding on the values did not require much tuning.

In each iteration, the camera parameters are computed using the 2D-to-3D correspondences. We initialize the parameters using the direct linear transform algorithm and refine it with the sparse bundle adjustment package [HZ04, LA09]. In order to obtain annota-

tions, we set up a simple system where the user can click on 2D images and click on the corresponding 3D points in the template mesh. Our system does not require the same vertices to be annotated in every image. The average number of points annotated for each image for both cats and horses was 29 points.

6. Results

Comparison with [CF13] Cashman et al. employ a low resolution control mesh on the order of less than 100 vertices which is then interpolated with Loop subdivision. In order to apply their method to ours, we simplified our template mesh with quadratic decimation until we reach around 150 vertices while retaining the key features of the template mesh as much as possible (shown in inset).

Since their method relies on silhouettes, we provide hand-segmented silhouettes to their algorithm along with the user-clicked points. We transferred the user-clicks from the full mesh to the simplified mesh by finding the closest 3D vertex in the simplified mesh for each labeled vertex in the full resolution mesh. We did not include points that did not have a close enough 3D vertex due to simplification. On average 24 points were labeled for their experiment and we use their default parameters.

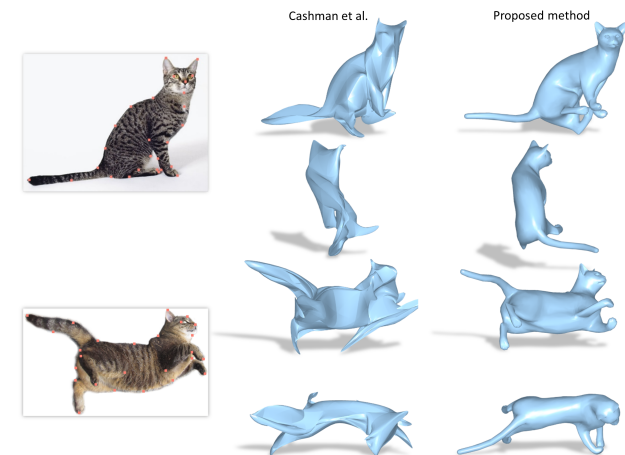


Figure 4: Comparison with [CF13]: the first column shows the user-clicked input images, the second column shows the result of [CF13] and the third column shows the result of our proposed method. Two views are shown for each image, one from the final estimated camera and another from an orthogonal view point. Our method is more robust to large deformations and retains the volume of the model. Note that silhouettes, along with the user-clicked points, are used to obtain the results for [CF13].

Figure 4 compares the results obtained with the method of [CF13] and our model. Two views are shown for each result, one from the estimated camera pose and another from an orthogonal viewpoint. As the authors in [CF13] point out, their method focuses on modeling shape and is not designed for highly articulated objects such as cats. Consequently, we can see that it has difficulties dealing with the wide range of poses present in our cat dataset. Regions such as limbs and tails especially lose their original shape.

Their method is based on surface deformation, which does not have a notion of volume. This causes flattening of the 3D models as can be seen in the orthogonal views. Since we guarantee worst-case distortion and orientation preserving deformation of the auxiliary mesh, our surface reconstructions are well behaved compared to [CF13]. Recall that silhouettes, along with the user-clicked points, are used to obtain the results for [CF13].

Qualitative evaluation The 3D models in Figure 1 were obtained using our proposed framework. We now evaluate the effectiveness of the proposed framework by comparing the results without any distortion bounds (i.e. removing Equation (6)) and with constant distortion bounds (i.e. fixing s_j to a constant). Qualitative results of this ablation study are shown in Figure 5. The first column shows input images along with their user-clicked points. The second column shows results obtained with no bounds, leaving just the ARAP energy, which we refer to as *Unbounded*. This is similar to the approach used in [KSES14], but with volumetric instead of surface deformation. The third column, *Uniform*, shows results obtained with a uniform bound, where the stiffness $1 + \epsilon + s_j$ is replaced with a single constant $c_j = 2$ for all faces. This is the deformation energy used in [KABL14] applied to 2D positional constraints. The constant was slowly increased from 1 to 2 in a manner similar to η in order to allow for increasing levels of deformation. Finally, in the last column we show results obtained with the proposed framework where the distortions are bounded with local stiffness.

First, notice the wide range of poses present in the images used; some are particularly challenging requiring large deformation from the template 3D mesh. In general, *Unbounded* concentrates high distortions near positional constraints causing unnatural stretching and deformation around limbs and faces. This is evident with horse legs in row 4 as *Unbounded* deforms them in an elastic manner. *Uniform* distributes the distortions, however, when the pose change from the template is significant, distortions spread out too much causing unrealistic results as seen in rows 2 and 3. The unnatural distortion of the faces is still a problem with *Uniform*. The proposed framework alleviates problems around the face and the horse limbs as it learns that those regions are more rigid. Please refer to the supplementary materials for comprehensive results of all cat and horse experiments.

We provide visualizations of the learned stiffness model in Figure 6 and 7. Figure 6 visualizes the learned stiffness values for cats and horses in various poses. We show the centroid of tetrahedra faces colored by their stiffness values in log scale. Blue indicates rigid regions while red indicates highly deformable regions. Recall that there is one stiffness model for each animal category. The level of deformations present in the input images are well reflected in the learned stiffness model. For cats, the torso is learned to be highly deformable allowing the animal to twist and curl, while keeping the skull and limbs more rigid. Similarly for horses, the neck, the regions connecting the limbs as well as the joints are learned to be deformable while keeping skull, limbs, and buttocks region rigid. The fact that the buttocks is considered rigid is anatomically consistent, since horses have a rigid spine making them suitable for riding [JD80].

We also present segmentation results using the learned stiffness values as another form of visualization in Figure 7. In order to ob-

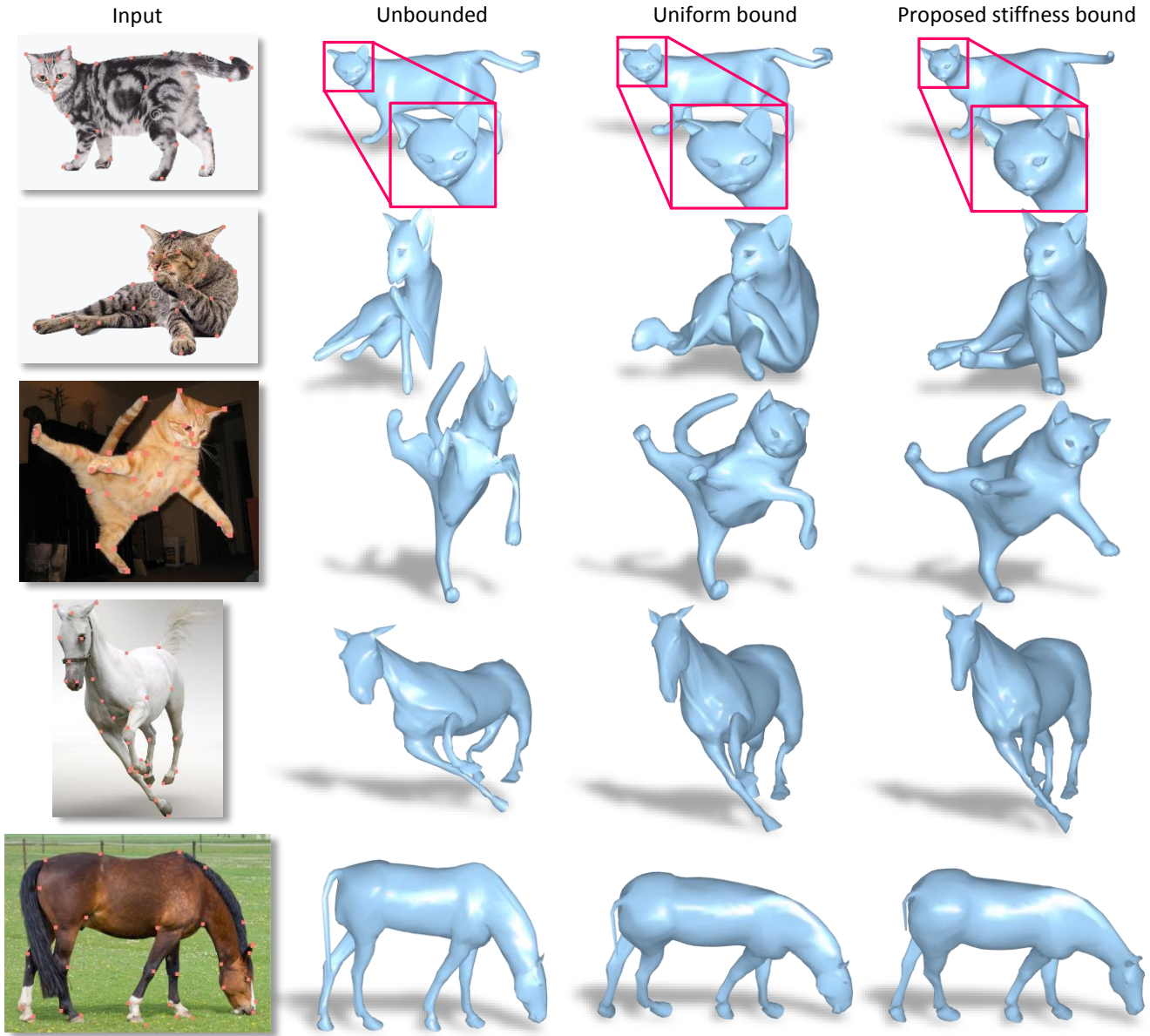


Figure 5: Comparison of proposed approach vs results with its key components removed. User-clicked input images (first column). Unbounded (second column) is the model without any bounds on the distortion leaving just the volumetric ARAP energy. Uniform (third column) is when the stiffness bounds (s_j in Equation (6)) are replaced by a single constant, which is the approach of [KABL14] applied to 2D positional constraints. The last column shows the results with our complete framework where the stiffness bounds and the deformations are jointly optimized. Without the animal-specific stiffness, distortions either spread out too much or concentrate around the positional constraints.

tain the segmentations, we first transferred the stiffness values from tetrahedra faces to vertices by taking the mean stiffness of faces a vertex is connected to. Then, we constructed a weighted graph on the vertices based on their connectivity, where the weights are set to be the sum of the Euclidean proximity and the similarity of the stiffness values. We apply normalized cuts to partition this graph and interpolate the result to the surface mesh vertices using the pa-

rameterization described in Section 3.1. We also show the segmentation results using just the Euclidean proximity as a comparison. Stiffness-based segmentation illustrates that regions which deform together as learned by our framework correspond to semantically reasonable parts.

The learned stiffness model can be used as a prior to solve for stiffness-aware deformations of new annotated images. Figure 8



Figure 6: Visualization of the learned stiffness values. Blue indicates rigid regions while red indicates highly deformable regions.

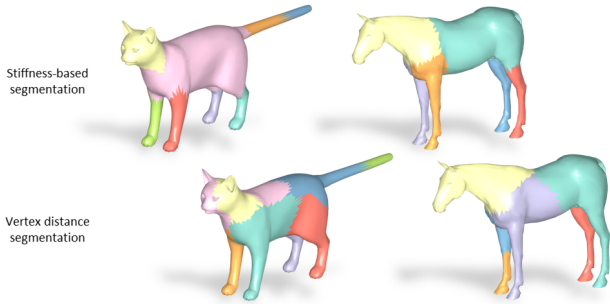


Figure 7: Another visualization of the learned stiffness by means of segmentation. Segmenting the template mesh using stiffness illustrates regions that deform together as learned by our framework. We see that they correspond to semantically reasonable segmentations. We show segmentation results based on vertex distance alone as a comparison.

shows the results of deforming the template to new input images via using the stiffness values learned from the previous experiment, i.e. the new images were not used to learn the stiffness. Similar to other experiments, we do warm start where the stiffness bounds are linearly increased from 1.01 to their actual value in 10 steps. The results are visually very similar to the results obtained when the stiffness was learned with those images along with the other 10 cat images. From this perspective, the joint optimization for the stiffness and the deformations using multiple images is the “training” (Figure 5), while the single-image optimization with a fixed stiffness prior is the “testing” (Figure 8).

Quantitative evaluation Lastly, we conduct an evaluation against the ground truth by using pictures of a rendered 3D model as the input to our framework. Specifically, we use the TOSCA dataset [BBK08], which has 11 vertex-aligned models of cats in various poses. We take the neutral pose (cat0) as the template and randomly project the other 10 models to produce images where the ground truth shape is known. We randomly sample 35 points and use them as the 3D-to-2D correspondences. In order to guarantee that these points are well distributed, we segment the model into 15 components and make sure that there is at least one point from each component. These components correspond to key parts such as the paws, limbs, left and right ears, tail base and tip, etc. We com-

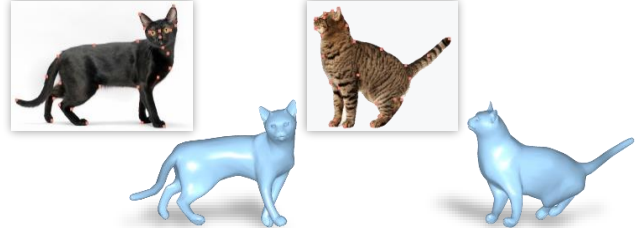


Figure 8: Deformation results using the learned stiffness from 10 cats as a fixed prior for new images.

Table 1: Quantitative evaluation against ground-truth. The Lower the better for all metrics.

Methods	Mean dist	Distortion error metric [YBS04]			
		Stretch	Edge	Area	Angle
Unbounded	0.291	1.01	0.156	0.216	0.159
Uniform	0.281	1.01	0.13	0.198	0.13
Proposed	0.287	0.996	0.105	0.181	0.085

pare the results of the No Bound, Uniform, and the proposed approach. Using this method, we produce two images from each ground truth model and conduct the experiment with 20 images.

We evaluate our method using several error metrics. First, we measure the distortions between the ground truth and the deformed models, which capture how natural the deformations are. We argue this is the most important measure since obtaining plausible deformations is the main goal of our algorithm. For this we use the stretch, edge-length, area, and angle distortion errors as defined in [YBS04] by comparing the corresponding triangles. Additionally, we report the mean Euclidean distance between the 3D vertices, which measures how close the surface of the deformed models are to the ground truth. While a low Euclidean distance is desirable for 3D reconstruction, we do not expect a close match everywhere due to ambiguities arising from a single view and sparse point constraints. In particular, Euclidean distance is not necessarily indicative of visual quality. We report the average error over all 20 input images. Before computing the error metrics, the deformed and ground truth models are aligned by a similarity transform. The results are shown in Table 1. As expected, all methods attain comparable mean Euclidean distance to ground truth, while our approach obtains substantially lower errors in distortion metrics. This demonstrates the advantage of learning stiffness from multiple images, yielding a more plausible deformation model.

Implementation details With an unoptimized MATLAB implementation, training with 10 images took 4 hours and testing a single image with a learned stiffness prior took ~30 minutes. We use YALMIP [LÖ4] for the SDP modeling and MOSEK as the solver [AA00]. Our biggest bottleneck is the SDP optimization due to many LMI constraints. Reducing the number of tets can significantly reduce the run-time.

7. Limitations and Future Work

Limitations of our current approach suggest directions for future work. One failure mode is due to a large pose difference between

the template and the target object, which may lead to an erroneous camera parameters estimate (e.g., local minima), as seen in row 5 of Figure 5. Here, the head of the horse in the image is lowered for grazing while the head of the horse template is upright causing a poor initialization of the camera estimate. Using a user-supplied estimate of the viewpoint or automatic viewpoint estimation methods like [TM15] are possible solutions.

Another pitfall is that some parts may be bent in an unnatural direction as seen around the left ankle of the horse in row 4 of Figure 5. An interesting future direction is to make the distortion bounds dependent on the orientation of the transformation. This would allow the framework to learn that certain parts only deform in certain directions.

Since we only have a single view for each target object, there is an inherent depth ambiguity, where many 3D shapes project to the same 2D observations. As such, some of our deformed models do not have the “right” 3D pose when seen from a different view. This is illustrated in our supplementary video that shows 360 degree views of the final models. One possibility is to combine our method with recent single image reconstruction approaches like [VCdAB14, CKTM15] that use a large image collection of the same object class to resolve the depth ambiguity.

Our method could also be enhanced to prevent surface intersections or reason about occlusion (e.g. if the point is labeled, it should be visible from the camera). Run-time is also an issue for adapting the stiffness model into a real-time posing application. This may be addressed by recent advancements in efficiently computing mappings with geometric bounds [KABL15].

8. Conclusion

Modifying 3D meshes to fit the pose and shape of objects in images is an effective way to produce 3D reconstruction of Internet images. In order to fit object pose naturally, it is essential to understand how an object can articulate and deform, especially for highly deformable and articulated objects like cats. In this paper we propose a method that can learn how an object class can deform and articulate from a set of user-clicked 2D images and a 3D template mesh. We do so by introducing a notion of local stiffness that controls how much each face of the mesh may distort. We jointly optimize for the deformed meshes and the stiffness values in an iterative algorithm where a convex optimization problem is solved in each iteration. Our experiments show that learning stiffness from multiple images produces more plausible 3D deformations. We hope this motivates further developments in the area of automatic point correspondence for non-rigid objects, enabling fully-automated 3D modeling of animals from 2D images in the near future.

Acknowledgements

We would like to thank Austin Myers for helpful comments and suggestions. This material is based upon work supported by the National Science Foundation under grant no. IIS-1526234, the Israel Binational Science Foundation, Grant No. 2010331 and the Israel Science Foundation Grants No. 1265/14.

References

- [AA00] ANDERSEN E. D., ANDERSEN K. D.: The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In *High performance optimization*. Springer, 2000, pp. 197–232. 8
- [AB15] AKHTER I., BLACK M. J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2015). 2
- [ACOL00] ALEXA M., COHEN-OR D., LEVIN D.: As-rigid-as-possible shape interpolation. In *ACM SIGGRAPH* (2000). 3
- [AKP*04] ANGUELOV D., KOLLER D., PANG H.-C., SRINIVASAN P., THRUN S.: Recovering articulated object models from 3d range data. In *UAI* (2004), pp. 18–26. 2
- [ASK*05] ANGUELOV D., SRINIVASAN P., KOLLER D., THRUN S., RODGERS J., DAVIS J.: SCAPE: Shape completion and animation of people. *ToG* 24, 3 (2005), 408–416. 2, 4
- [BBK07] BRONSTEIN A. M., BRONSTEIN M. M., KIMMEL R.: Calculus of nonrigid surfaces for geometry and texture manipulation. *IEEE Transactions on Visualization and Computer Graphics* 13, 5 (2007). 5
- [BBK08] BRONSTEIN A. M., BRONSTEIN M. M., KIMMEL R.: *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008. 8
- [BPGK06] BOTSCH M., PAULY M., GROSS M. H., KOBBELT L.: Prmo: coupled prisms for intuitive surface modeling. In *Symposium on Geometry Processing* (2006), pp. 11–20. 1, 2
- [BS08] BOTSCH M., SORKINE O.: On linear variational surface deformation methods. *Visualization and Computer Graphics, IEEE Transactions on* 14, 1 (2008), 213–230. 2
- [BTG*12] BALLAN L., TANEJA A., GALL J., GOOL L. V., POLLEFEYS M.: Motion capture of hands in action using discriminative salient points. In *ECCV* (October 2012). 2
- [BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH* (1999), pp. 187–194. 2
- [BV04] BOYD S., VANDENBERGHE L.: *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. 5
- [CF] CASHMAN T. J., FITZGIBBON A. W.: Forms: Flexible object reconstruction from multiple silhouettes. <http://forms.codeplex.com/>. 5
- [CF13] CASHMAN T. J., FITZGIBBON A. W.: What shape are dolphins? building 3D morphable models from 2D images. *IEEE Trans. Pattern Anal. Mach. Intell* 35, 1 (2013), 232–244. 2, 5, 6
- [CKC10] CHEN Y., KIM T.-K., CIPOLLA R.: Inferring 3D shapes and deformations from single views. In *ECCV* (2010). 2
- [CKTM15] CARREIRA J., KAR A., TULSIANI S., MALIK J.: Virtual view networks for object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)* (2015), IEEE. 9
- [dAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. *ACM SIGGRAPH* 27, 3 (2008). 3
- [dATTS08] DE AGUIAR E., THEOBALT C., THRUN S., SEIDEL H.-P.: Automatic conversion of mesh animations into skeleton-based animations. *Computer Graphics Forum (Proc. Eurographics EG'08)* 27, 2 (2008), 389–397. 2
- [DSP06] DER K. G., SUMNER R. W., POPOVIĆ J.: Inverse kinematics for reduced deformable models. In *ACM Transactions on Graphics* (2006), ACM. 2
- [FP10] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell* 32, 8 (2010), 1362–1376. 1
- [GWBB09] GUAN P., WEISS A., BALAN A. O., BLACK M. J.: Estimating human shape and pose from a single image. In *ICCV* (2009), IEEE, pp. 1381–1388. 2

- [HSS*09] HASLER N., STOLL C., SUNKEL M., ROSENHAHN B., SEIDEL H.-P.: A statistical model of human pose and body shape. *Comput. Graph. Forum* 28, 2 (2009), 337–346. 2
- [hTgL10] HUA TAN G., 0001 W. C., GANG LIU L.: Image driven shape deformation using styles. *Journal of Zhejiang University - Science C* 11, 1 (2010), 27–35. 2
- [HZ04] HARTLEY R. I., ZISSERMAN A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004. 2, 3, 5
- [IKH*11] IZADI S., KIM D., HILLIGES O., MOLYNEAUX D., NEWCOMBE R., KOHLI P., SHOTTON J., HODGES S., FREEMAN D., DAVISON A., ET AL.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST* (2011), ACM. 1
- [JD80] JEFFCOTT L. B., DALIN G.: Natural rigidity of the horse's backbone. *Equine veterinary journal* 12, 3 (1980), 101–108. 6
- [KABL14] KOVALSKY S. Z., AIGERMAN N., BASRI R., LIPMAN Y.: Controlling singular values with semidefinite programming. *ACM SIGGRAPH* 33, 4 (2014), 68. 1, 4, 5, 6, 7
- [KABL15] KOVALSKY S. Z., AIGERMAN N., BASRI R., LIPMAN Y.: Large-scale bounded distortion mappings. *ACM Transactions on Graphics (Special Issue of SIGGRAPH Asia)* 34, 6 (Oct. 2015), 191:1–191:10. 9
- [KSES14] KHOLGADE N., SIMON T., EFROS A. A., SHEIKH Y.: 3D object manipulation in a single photograph using stock 3D models. *ACM SIGGRAPH* 33, 4 (2014), 127. 2, 6
- [KSvdP09] KRAEVOY V., SHEFFER A., VAN DE PANNE M.: Modeling from contour drawings. In *SBM* (2009), Eurographics Association. 2
- [LÖ4] LÖFBERG J.: YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference* (2004). URL: <http://users.isy.liu.se/johanl/yalmip>. 8
- [LA09] LOURAKIS M. I., ARGYROS A. A.: Sba: A software package for generic sparse bundle adjustment. *TOMS* (2009). 5
- [LD14] LE B. H., DENG Z.: Robust and accurate skeletal rigging from mesh sequences. *ACM Transactions on Graphics* 33, 4 (2014), 84. 2
- [Lev98] LEVIN D.: The approximation power of moving least-squares. *Math. Comput* 67, 224 (1998). 3
- [Lip12] LIPMAN Y.: Bounded distortion mapping spaces for triangular meshes. *ACM SIGGRAPH* 31, 4 (2012), 108. 4
- [LSCO*04] LIPMAN Y., SORKINE O., COHEN-OR D., LEVIN D., ROSSI C., SEIDEL H.-P.: Differential coordinates for interactive mesh editing. In *Shape Modeling Applications, 2004. Proceedings* (2004), IEEE, pp. 181–190. 2
- [MTGG11] MARTIN S., THOMASZEWSKI B., GRINSPUN E., GROSS M.: Example-based elastic materials. *ACM Transactions on Graphics* (2011). 2
- [PJS06] POPA T., JULIUS D., SHEFFER A.: Material-aware mesh deformations. In *Shape Modeling International* (2006). 1, 2
- [SA07] SORKINE O., ALEXA M.: As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing* (2007), Eurographics Association. 2, 3
- [Si15] SI H.: Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Trans. Math. Softw* 41, 2 (2015), 11. 5
- [SP04] SUMNER R. W., POPOVIC J.: Deformation transfer for triangle meshes. *ACM SIGGRAPH* 23, 3 (2004), 399–405. 2
- [SSP07] SUMNER R. W., SCHMID J., PAULY M.: Embedded deformation for shape manipulation. *ACM Transactions on Graphics* 26, 3 (2007), 80. 2
- [SY07] SCHAEFER S., YUKSEL C.: Example-based skeleton extraction. In *SGP07: Eurographics Symposium on Geometry Processing* (2007), pp. 153–162. 2
- [SZGP05] SUMNER R. W., ZWICKER M., GOTSMAN C., POPOVIC J.: Mesh-based inverse kinematics. *ACM SIGGRAPH* 24, 3 (2005), 488–495. 2
- [TM15] TULSIANI S., MALIK J.: Viewpoints and keypoints. In *Computer Vision and Pattern Recognition (CVPR)* (2015). 9
- [TPBF87] TERZOPOULOS D., PLATT J., BARR A., FLEISCHER K.: Elastically deformable models. In *ACM SIGGRAPH* (1987), ACM, pp. 205–214. 2
- [TSR*14a] TAYLOR J., STEBBING R., RAMAKRISHNA V., KESKIN C., SHOTTON J., IZADI S., HERTZMANN A., FITZGIBBON A.: User-specific hand modeling from monocular depth sequences. In *Computer Vision and Pattern Recognition (CVPR)* (2014), IEEE, pp. 644–651. 2
- [TSR*14b] TAYLOR J., STEBBING R. V., RAMAKRISHNA V., KESKIN C., SHOTTON J., IZADI S., HERTZMANN A., FITZGIBBON A. W.: User-specific hand modeling from monocular depth sequences. In *Computer Vision and Pattern Recognition (CVPR)* (2014). 2
- [VCdAB14] VICENTE S., CARREIRA J., DE AGAPITO L., BATISTA J.: Reconstructing PASCAL VOC. In *Computer Vision and Pattern Recognition (CVPR)* (2014). 9
- [VdA13] VICENTE S., DE AGAPITO L.: Balloon shapes: Reconstructing and deforming objects with volume from images. In *3DV* (2013), IEEE. 2
- [YBS04] YOSHIZAWA S., BELYAEV A., SEIDEL H.-P.: A fast and simple stretch-minimizing mesh parameterization. In *Shape Modeling Applications, 2004. Proceedings* (2004), IEEE, pp. 200–208. 8
- [YP08] YAN J. Y., POLLEFEYS M.: A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE Trans. Pattern Anal. Mach. Intell* (2008). 4
- [YY14] YE M., YANG R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Computer Vision and Pattern Recognition (CVPR)* (2014), IEEE. 2
- [ZHS*05] ZHOU K., HUANG J., SNYDER J., LIU X., BAO H., GUO B., SHUM H.-Y.: Large mesh deformation using the volumetric graph laplacian. *ACM Transactions on Graphics* 24, 3 (2005), 496–503. 2, 3